NUMERIEKE ANALYSE III

NUMERIEKE BEHANDELING VAN DIFFERENTIAALVERGELIJKINGEN

A. van der Sluis

1975

Mathematisch Instituut
der Rijksuniversiteit
Utrecht.

§45. Recurrente betrekkingen.

(45.1) Bij de numerieke behandeling van differentiaalvergelijkingen loopt men voortdurend tegen recurrente betrekkingen aan. Daarom bespreken we nu eerst enkele eigenschappen ervan (een deel hiervan vindt men al in TA (12.9) t/m (12.16)).

(45.2) Onder een lineaire k+1-terms recurrente betrekking voor een rij $\{a_i\}_{i\geq 0}$ verstaat men een relatie

(45.3) $a_{i+k} + c_{1i}a_{i+k-1} + \dots + c_{ki}a_i = g_{i+k}$, $i = 0,1,2,\dots$.

met gegeven c_{ji}, $j = 1,\dots,k$; $i = 0,1,2,\dots$ Men noemt zo'n rij $\{a_i\}$ ook wel een oplossing van (45.3). Een oplossing is blijkbaar bepaald door (willekeurig voor te schrijven) beginwaarden $a_0,a_1,\dots,a_{k-1}$.

(45.4) Een bijzonder geval is dat de c_{ji} niet van i afhangen, dus b.v.

$a_{i+2} + 6a_{i+1} + 2a_i = g_{i+2}$, $i = 0,1,2,\dots$

in welk geval men spreekt van een reeurrente betrekking met constante coëfficienten.

(45.5) Als $g_j = 0$ voor alle j noemt men de recurrente betrekking homogeen.

Men verifieert onmiddellijk:

(45.6) Als (45.3) homogeen is vormen de oplossingen een k-dimensionale lineaire ruimte d.w.z. dat als $\{a_{1i}\}$ en $\{a_{2i}\}$ twee rijen zijn die aan (45.3) met alle $g_i=0$ voldoen, dan voldoet ook elke lineaire combinatie $\{pa_{1i}+qa_{2i}\}$; en er zijn rijen $\{a_{1i}\},\dots\{a_{ki}\}$ zodat elke oplossing van de homogene vergelijking een lineaire combinatie hiervan is (neem nl. maar voor $\{a_{1i}\}$ de rij waarvan de eerste k termen luiden 1,0,0,...,0, voor $\{a_{2i}\}$ de rij die begint met 0,1,0,...,0 etc.).

(45.7) Een willekeurige oplossing van de inhomogene vergelijking (45.3) wordt verkregen door een vaste oplossing van (45,3) te vermeerderen met een geschikte oplossing van de homogene recursie die bij (45.3) past. Blijkbaar vormen alle oplossingen van (45,3) een k-dimensionale lineaire variëteit.

(45.8) Van een lineaire homogene recursie met constante coëfficienten laten de oplossingen zich op zeer eenvoudige wijze voorstellen:

Stelling. Zij

(45.9) $a_{i+k} + c_1a_{i+k-1} + \dots + c_ka_i = 0$

een recurrente betrekking met constante coëfficienten.

Als de <u>karakteristieke vergelijking</u> van deze recursie:

(45.10) $f(x)=x^k + c_1x^{k-1} + \ldots + c_k = 0$

uitsluitend enkelvoudige wortels $\lambda_1,\ldots,\lambda_k$ heeft dan vormen de rijen $\{\lambda_j^i\}_{i\geq 0}$ een basis van de oplossingsruimte, d.w.z. voor een willekeurige oplossing $\{a_i\}$ bestaan constanten $q_1,\ldots,q_k$ zodat

(45.11) $a_i = q_1\lambda_1^i + \ldots + q_k\lambda_k^i.$

Als, algemener, de karakteristieke vergelijking de verschillende wortels $\lambda_1,\ldots,\lambda_p$ heeft met multipliciteiten $\mu_1,\ldots,\mu_p$ dan vormen de rijen $\{i^m\lambda_j^i\}_{i\geq 0}$, $m = 0,1,\ldots,\mu_{j-1}$; $j = 1,\ldots,p$, een basis van de oplossingsruimte, d.w.z.

(45.12) $a_i = q_1(i)\lambda_1^i + \ldots + q_p(i)\lambda_p^i$

met q_j een veelterm van graad μ_{j-1}.

<u>Bewijs</u>. Zij $i_m = i(i-1)\ldots(i-m+1)$ voor m geheel > 0, $i_0 = 1$, voor alle i. Wegens $f^{(m)}(\lambda_1) = 0$ voor $m = 0,1,\ldots,\mu_1-1$ geldt $\frac{d^m}{dx^m}(x^if(x))_{x=\lambda_1} = 0$ (regel van Leibniz voor het herhaald differentieren van een produkt). Dus

$$[(i+k)_mx^{i+k} + c_1(i+k-1)_mx^{i+k-1} + \ldots + c_ki_mx^i]x^{-m} = 0.$$

Hieruit volgt meteen dat $\{i_m\lambda_1^i\}_{i\geq 0}$, $m = 0,1,\ldots,\mu_1-1$ een oplossing is van (45.9). Hetzelfde geldt voor $\lambda_2,\ldots,\lambda_p$.

We krijgen op deze manier precies k oplossingen, en zij zullen dus een basis van de oplossingsruimte vormen als ze lineair onafhankelijk zijn. Dat dit laatste inderdaad het geval is laten we eenvoudigheidshalve aan een speciaal geval zien; het algemene geval gaat echter analoog. Zij $p=3$, $\mu_1=1$, $\mu_2=2$, $\mu_3=3$. We noteren van elk der genoemde oplossingen de eerste 6 termen.

1	λ_1	λ_1^2	λ_1^3	λ_1^4	λ_1^5
1	λ_2	λ_2^2	λ_2^3	λ_2^4	λ_2^5
0	1	$2\lambda_2$	$3\lambda_2^2$	$4\lambda_2^3$	$5\lambda_2^4$
1	λ_3	λ_3^2	λ_3^3	λ_3^4	λ_3^5
0	1	$2\lambda_3$	$3\lambda_3^2$	$4\lambda_3^3$	$5\lambda_3^4$
0	0	2.1	$3.2\lambda_3$	$4.3\lambda_3^2$	$5.4\lambda_3^3$

We beweren dat reeds deze 6 rijtjes lin-onafhankelijk zijn en daarmee de hele oplossingen. De onafhankelijkheid van deze rijtjes volgt uit die van de kolommen. Stel derhalve dat $r_1 \times$ 1e kolom + + $r_2 \times$ 2e kolom + ... = 0. Zij $g(x) = r_1 + r_2x + \ldots + r_6x^5$. Dan geldt blijkbaar $g(\lambda_1) = g(\lambda_2) = g'(\lambda_2) = g(\lambda_3) = g'(\lambda_3) = g''(\lambda_3) = 0$, zodat g deelbaar is door $(x-\lambda_1)(x-\lambda_2)^2(x-\lambda_3)^3$ en dat kan wegens de graad van g alleen als g identiek 0 is, dus $r_1 = r_2 = \ldots = r_6 = 0$.

Rest nog te bewijzen dat ook $\{i^m\lambda_j^i\}_{i\geq 0}$, $m = 0,1,\ldots,\mu_j-1$; $j = 1,\ldots,p$ een basis vormen. Dit volgt uit het feit dat ook dit k oplossingen zijn en dat de eerder genoemde basis hieruit door lineaire combinatie ontstaat, zulks omdat $\{i_m\}$ lin. combinatie is van $\{i^m\},\{i^{m-1}\}, \ldots,\{i^0\}$.

(45.13) Men kan een (k+1)-terms recurrente betrekking vertalen in een twee terms vector recurrente betrekking, hetgeen voor het vervolg heel nuttig zal blijken. Definieer nl. rijen vectoren $\{v_i\}$ en $\{b_i\}$ en matrices A_i als volgt:

(45.14) $$v_i = \begin{bmatrix} a_i \\ \vdots \\ a_{i+k-1} \end{bmatrix}, \quad b_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ g_{i+k} \end{bmatrix}, \quad A_i = \begin{bmatrix} 0 & 1 & 0 \cdots\cdots 0 \\ 0 & 0 & 1 \cdots\cdots 0 \\ \vdots & \ddots & \ddots \quad \vdots \\ 0 & 0 & \ddots \quad 1 \\ -c_{ki} & -c_{k-1,i} & \cdots \quad -c_{1,i} \end{bmatrix}$$

Dan wordt (45.3)

(45.15) $v_{i+1} = A_i v_i + b_i$, $i = 0,1,2,\ldots$

(45.16) In het bijzondere geval van een homogene vergelijking met constante coëfficienten geldt

(45.17) $v_{i+1} = Av_i$

zodat

(45.18) $v_i = A^i v_o$.

Uit (45.8) volgt:

(45.19) De oplossingen van (45.17) zijn dan en slechts dan begrensd als alle wortels binnen of op de eenheidscirkel liggen en bovendien de wortels op de eenheidscirkel enkelvoudig zijn.

(45.20) Opgave. Ga na dat uit (41.5) volgt dat voor een willekeurige matrix A geldt dat $\{A^i\}$ dan en slechts dan begrensd is als alle eigenwaarden binnen of op de eenheidscirkel liggen en bovendien de eigenwaarden die op de eenheidscirkel liggen uitsluitend 1 × 1 Jordankastjes hebben.

(45.21) Opgave. Dat in (45.17) alle wortels op de eenheidscirkel enkelvoudig moeten zijn terwijl dit in (45.20) niet hoeft zit hem hierin dat in A volgens (45.14) bij elke eigenwaarde, of deze nu enkelvoudig is of niet, slechts één eigenvector behoort, en dus ook slechts één Jordankast. Ga dit na.

(45.22) We beschouwen weer (45.15), waarbij A,v_i en b_i niet noodzakelijk de gedaante uit (45.14) hoeven te hebben.

Er geldt $v_{n+1} = b_n + A_n b_{n-1} + \dots + A_n A_{n-1} \cdots A_1 b_0 + A_n A_{n-1} \cdots A_0 v_0$, zodat

(45.23) $$\|v_{n+1}\| \leq \max_{0 \leq i \leq n} \|A_n A_{n-1} \cdots A_i\| \left(\|v_0\| + \sum_{i=0}^{n} \|b_i\|\right).$$

Ons zal de situatie interesseren dat $A_i = A + Q_i$.

Hiervoor geldt

(45.24) Stelling. Als $\|A^i\| \leq M$, $M \geq 1$, en $\|Q_i\| \leq K$ voor $i \leq n$ dan geldt

$$\left\| \prod_{l=i+1}^{n} (A+Q_l) \right\| \leq M(1+MK)^{n-i}.$$

Bewijs. $\prod_{i+1}^{n} (A+Q_l)$ bestaat uit termen van de gedaante $A \dots A \; Q \; A \dots A \; Q \; A \dots A$ waarbij sommige der rijtjes A leeg kunnen zijn. Een term met j factoren Q heeft hoogstens $j+1$ rijtjes A en is in norm dus $\leq M^{j+1}K^j$. Er zijn $\binom{n-i}{j}$ van die termen. Dus

$$\left\| \prod_{l=i+1}^{n} (A+Q_l) \right\| \leq \sum_{j=0}^{n-i} \binom{n-i}{j} M^{j+1} K^j.$$

§46. Probleemstelling beginwaarde probleem. Afspraken.

(46.1) We beschouwen de gewone eerste orde diff. vergelijkingen.

(46.2) $$\frac{dx}{dt} = f(t,x)$$

voor $t \in I = [t_0, t_0+T]$ en $x \in \mathbb{R}^n$.

Zoals bekend zijn differentiaalvergelijkingen van het type

(46.3) $$\frac{d^n u}{dt^n} = g(t,u,u',\dots,u^{(n-1)})$$

op de gedaante (46.2) te brengen door nl. te definiëren

$$f(t,x) = (\xi_1, \xi_2, \dots, \xi_{n-1}, g(t,\xi_0,\xi_1,\dots,\xi_{n-1}))^T$$

als $x(t) = (\xi_0(t), \xi_1(t), \dots, \xi_{n-1}(t))$. (Merk op dat dan geldt $u^{(k)}(t) = \xi_k(t)$).

Dit geldt dus in het bijzonder ook voor diff. vergelijkingen van het type

(46.4) $$u^{(n)} + a_1(t)u^{(n-1)} + \dots + a_n(t)u = v(t).$$

(46.5) Gewoonlijk heeft een dvgl. vele oplossingen en moet (kan) men dus verdere condities stellen aan de oplossing. Bekend is het zgn. beginwaarde probleem waarbij men eist dat $x(t_0) = x_0$, x_0 een gegeven vector.

(46.6) Een voldoende voorwaarde opdat het beginwaarde probleem althans op een omgeving van t_0 precies één oplossing heeft is dat op een gebied $\Omega \subset I \times \mathbb{R}^n$, dat het punt (t_0, x_0) bevat, geldt:

a) f continu en

b) f Lipschitz in de tweede variabele;

d.w.z. dat er een constante L bestaat zodat voor alle t,x en $\tilde{x}$ waarvoor (t,x) en $(t,\tilde{x}) \in \Omega$:

(46.7) $$\|f(t,x) - f(t,\tilde{x})\| \leq L\|x-\tilde{x}\|.$$

Deze voorwaarde is niet voldoende om te garanderen dat de bedoelde oplossing ook op geheel I gedefinieerd is. Beschouw bv. het beginwaarde probleem $\frac{dx}{dt} = x^2$, $x(0) = 1$, met als oplossing $x(t) = \frac{1}{1-t}$. Men ziet dat $f(t,x) = x^2$ op een willekeurig groot gebied Lipschitz is, maar dat desondanks de oplossing slechts tot $t=1$ bestaat.

(46.8) We zullen verder aannemen dat het bewuste beginwaarde probleem inderdaad een unieke oplossing x^* op I heeft, en dat het gebied Ω waarvan in (46.6) sprake was, de grafiek van x^* bevat, dus de verzameling punten $(t,x^*(t))$ met $t \in I$.

(46.9) Zij verder $t_n=t_o+nh$, $h = T/N$, N een natuurlijk getal. Men noemt h de <u>stapgrootte</u>.

Met x_n^* resp. x_n geven we de waarde aan van de exacte resp. een benaderde oplossing in het punt t_n.

(46.10) We zullen in het vervolg methoden bezien waarmee benaderde oplossingen x_n in de punten t_n kunnen worden bepaald. De problemen zijn daarbij ruwweg de volgende:

1. hoe groot is de fout $x_n-x_n^*$ afhankelijk van h en $nh \in I$ vast. Geldt $x_n-x_n^* \to 0$ voor nh vast (d.w.z. voor een vaste waarde van t) en $h \to 0$? (convergentie).
2. Hoe stabiel is de methode b.v. t.o.v. afrondfouten of kleine verstoringen in $x(t_o)$? (het begrip stabiliteit zal expliciet gedefinieerd worden en een belangrijk hulpmiddel zijn ter beantwoording van vraag 1.).
3. Is de methode wel efficiënt? (Het zal blijken dat niet iedere methode met succes op elke differentiaalvergelijking toepasbaar is, stijve differentiaalvergelijkingen).

§47. Enkele eenvoudige oplosmethoden.

(47.1) Een aantal bekende oplosmethoden berusten op de opmerking dat voor de ware oplossing geldt

(47.2) $$x_n^* = x_{n-p}^* + \int_{t_{n-p}}^{t_n} f(s,x^*(s))ds.$$

Laten nu $x_0, x_1, \ldots, x_{n-1}$ al bepaald zijn.
Zij $L^*(s)$ het Lagrange interpolatiepolynoom van de functie $s \to f(s,x^*(s))$, op een aantal der punten $t_0,\ldots,t_{n-1}$.
Men benadert nu (47.2) door

(47.3) $$x_n = x_{n-p} + \int_{t_{n-p}}^{t_n} L(s)ds$$

waarin L een interpolatiepolynoom met dezelfde steunpunten als L^* is maar met functiewaarden $f(t_i,x_i)$ i.p.v. $f(t_i,x_i^*)$ voor L^*.
Enkele eenvoudige gevallen:

(47.4) $$x_n = x_{n-1} + hf(t_{n-1},x_{n-1}).$$ (Methode van Euler).

Op $[t_{n-1},t_n]$ wordt $f(s,x^*(s))$ benaderd door $L^*(s) = f(t_{n-1},x^*_{n-1})$.

(47.5) $$x_n = x_{n-1} + \tfrac{3}{2}hf(t_{n-1},x_{n-1})-\tfrac{1}{2}hf(t_{n-2},x_{n-2}).$$

Op $[t_{n-2},t_n]$ wordt $f(s,x^*(s))$ benaderd door het lineaire interpolatiepolynoom L^* op de punten t_{n-2}, t_{n-1}. Wordt vervolgens geïntegreerd van t_{n-1} naar t_n.

(47.6) Bovenstaande formules hebben gemeen dat zij op extrapolatie gebaseerd zijn. Men noemt ze <u>expliciete</u> formules. <u>Impliciete</u> formules verkrijgt men door in de benadering van $f(s,x(s))$ in (47.3) het punt t_n (waar f nog onbekend is) mee te nemen.

Enkele eenvoudige gevallen:

(47.7) $$x_n = x_{n-1} + \frac{h}{2}[f(t_n,x_n) + f(t_{n-1},x_{n-1})].$$ (trapeziumregel).

Op $[t_{n-1},t_n]$ wordt f benaderd door het interpolatiepolynoom in de punten t_{n-1} en t_n. Dit wordt vervolgens van t_{n-1} naar t_n geïntegreerd.

(47.8) $$x_n = x_{n-1} + \frac{h}{12}[5f(t_n,x_n) + 8f(t_{n-1},x_{n-1}) - f(t_{n-2},x_{n-2})].$$

Ga zelf na hoe deze methode verkregen kan worden.

(47.9) Om bij een impliciete formule x_n te berekenen moet i.h.a. een niet-lineaire vergelijking (of stelsel niet-lineaire vergelijkingen) voor x_n worden opgelost. Dit gebeurt veelal met successieve substitutie (zie 14.9). Men kan aantonen dat dit proces in vele gevallen convergeert met convergentiefactor hL, mits $hL < 1$. (Merk op dat $hL < 1$ geen voldoende voorwaarde voor convergentie is).

(47.10) De verkregen benaderingsmethoden hebben alle de vorm

(47.11) $$x_n+\alpha_1 x_{n-1} + \ldots + \alpha_k x_{n-k} = h[\beta_0 f(t_n x_n)+\ldots+\beta_k f(t_{n-k},x_{n-k})].$$

Zo'n algorithme noemt men een k-steps methode, algemeen een multistepmethode of kortweg een multistep voor $k \geq 2$. Voor $k=1$ spreekt men van een eenstapsmethode. In feite is (47.11) een i.h.a. niet-lineaire recurrente betrekking.

(47.12) Er is één opvallend verschil tussen multistep methoden en eenstapsmethoden:
indien $x(t_o) = x_o$ gegeven is kan men direct een eenstapsmethode toepassen, echter niet een multistep. Daarvoor moet men eerst op een andere manier (bv. m.b.v. een eenstapsmethode) de startwaarden $x_1, x_2, \ldots, x_{k-1}$ bepalen. Men spreekt van het starten van de multistep en van de startprocedure ter bepaling van de startwaarden.
We komen hier later op terug.

§48. Algemene theorie van multistepmethoden.

(48.1) Bij het bestuderen van methodes zoals (47.11) interesseert ons natuurlijk de approximatiefout $\max\limits_{n \leq N} |x_n - x_n^*|$, ook wel globale discretiefout genoemd.
Hierbij is uiteraard van belang
a) hoe goed (47.11) de differentiaalvergelijking benadert
b) of (47.11) in geval $\beta_o \neq 0$ wel oplosbaar is naar x_n
c) hoe stabiel (47.11) is t.a.v. kleine verstoringen; deze zúllen vrijwel zeker ontstaan door fouten in de beginwaarden $x_o, \ldots, x_{k-1}$ en doordat x_n niet precies aan (47.11) voldoet.

(48.2) Met a) bedoelen we hoe goed een oplossing van de differentiaalvergelijking aan (47.11) voldoet. Hierbij tekenenwe aan dat, omdat in (47.11) f optreedt vermenigvuldigd met h, men eigenlijk moet verwachten dat (47.11) eerst links en rechts gedeeld moet worden door h om een approximatie voor de differentiaalvergelijking te kunnen zijn. Dit is bv. ook heel duidelijk uit (47.4) waar $(x_n - x_{n-1})/h$ een benadering is voor de afgeleide van x^* in t_{n-1}.

(48.3) Definitie. Zij
$$h\delta_n = x_n^* + \alpha_1 x_{n-1}^* + \ldots + \alpha_k x_{n-k}^* - h[\beta_o f(t_n, x_n^*) + \ldots + \beta_k f(t_{n-k}, x_{n-k}^*)].$$
Dan heet δ_n de locale discretisatiefout (ook wel truncation error).

(48.4) Definitie. De multistep (47.11) heet consistent als $\delta_n \to 0$ voor $h \to 0$ uniform in n mits $nh \leq T$.

(48.5) Definitie. De multistep (47.11) heet consistent en van de orde p (kortweg van de orde p) voor een zekere differentiaalvergelijking als p het grootste getal is waarvoor

$\delta_n = O(h^p)$ voor $h \to 0$ uniform in n,

mits $nh \leq T$

(48.6) Voorbeeld van wat er kan gebeuren bij een niet consistente methode: diff. verg.: $\frac{dx}{dt} = 1$, $t \in \mathbb{R}$, $x^*(0) = 0$ (dus ware oplossing $x^*(t) = t$). Multistep: $x_n - x_{n-1} = h[4f(t_{n-1},x_{n-1}) + f(t_{n-2},x_{n-2})]$. Voor de gegeven dvgl. wordt dit

$$x_n - x_{n-1} = 5h$$

en er zijn voor deze bijzondere f geen startmoeilijkheden: met $x_0 = 0$ is de oplossing geheel bepaald en wel $x_n = 5nh$, zodat men voor willekeurig kleine h met $nh = t$ vindt $x_n = 5t$, hetgeen ver afligt van de $x^*(t) = t$. De multistep is dan ook niet consistent: $h\delta_n = x_n^* - x_{n-1}^* - 5h = h - 5h = -4h$ wegens $x^*(t) = t$, zodat $\delta_n = -4$, hetgeen niet naar 0 gaat voor $h \to 0$.

(48.7) Voorbeeld hoe de orde te bepalen: differentiaalvergelijking: $\frac{dx}{dt} = f(t,x)$, $t \in I$, $x(h) = x_0$ met unieke oplossing $x^*(t)$, $t \in I$.
Multistepmethode: $x_n - \frac{4}{3}x_{n-1} + \frac{1}{3}x_{n-2} = \frac{2}{3}hf(t_n,x_n)$.
Neem aan $f \in C^{(2)}$; dan $x^* \in C^{(3)}$.

$$x_n^* - \frac{4}{3}x_{n-1}^* + \frac{1}{3}x_{n-2}^* - \frac{2}{3}hf(t_n,x_n^*) = x_n^* - \frac{4}{3}x_{n-1}^* + \frac{1}{3}x_{n-2}^* - \frac{2}{3}hx_n^{*\prime} =$$

(Taylor rond $t = t_{n-1}$)

$$= (x_{n-1}^* + hx_{n-1}^{*\prime} + \frac{h^2}{2}x_{n-1}^{*\prime\prime} + O(h^3)) - \frac{4}{3}x_{n-1}^* + \frac{1}{3}(x_{n-1}^* - hx_{n-1}^{*\prime} + \frac{h^2}{2}x_{n-1}^{*\prime\prime} + O(h^3)) - \frac{2}{3}h(x_{n-1}^{*\prime} + hx_{n-1}^{*\prime\prime} + O(h^2)) =$$

$$= (1 - \frac{4}{3} + \frac{1}{3})x_{n-1}^* + h(1 - \frac{1}{3} - \frac{2}{3})x_{n-1}^{*\prime} + h^2(\frac{1}{2} + \frac{1}{6} - \frac{2}{3})x_{n-1}^{*\prime\prime} + O(h^3) = O(h^3).$$

Dus is de multistep consistent en van orde 2 mits $f \in C^{(3)}$. Ga zelf na wat de orde is indien $f \in C^{(2)}$, $f \in C^{(1)}$.

(48.8) Opmerking. De orde van een multistep hangt kennelijk van f af. In de praktijk maken we ons van dit probleem af door f steeds "voldoende vaak continu differentieerbaar" te veronderstellen wanneer men zegt dat de orde van een multistep p is.

(48.9) Opgave. Bepaal de orde van (47.4), (47.5), (47.7) en (47.8).

(48.10) Zij ρ het polynoom

$$\rho(z) = z^k + \alpha_1 z^{k-1} + \ldots + \alpha_k$$

en σ het polynoom

$$\sigma(z) = \beta_o z^k + \beta_1 z^{k-1} + \ldots + \beta_k .$$

(48.11) Stelling. Noodzakelijk en voldoende voor consistentie is: $\rho(1) = 0$, $\rho'(1) = \sigma(1)$.

Bewijs. Bedenk dat $x^* \in C^1(I)$ en bepaal δ_n door rond $t = t_{n-k}$ in Taylor te ontwikkelen.

(48.12) Consistentie is niet toereikend voor plezierig gedrag van een methode. Beschouw bv. eens de multistep $x_n - 6x_{n-1} + 5x_{n-2} = h[\beta_o f(t_n, x_n) + \beta_1 f(t_{n-1}, x_{n-1})]$. Voor geschikte β_o en β_1 is deze methode zeker consistent. Bekijk echter nu eens de dvgl. $\frac{dx}{dt} = 0$, $t \in \mathbb{R}$, $x(0) = 1$.

Voor de gegeven dvgl. wordt x_n uit $x_o = x^*(t_o) = 1$ en x_1 bepaald door de recursie

$$x_n - 6x_{n-1} + 5x_{n-2} = 0.$$

Neem aan $x_1 = 1$. Dan geldt blijkbaar

$$x_n = +1.$$

Neem nu eens $x_1 = 1 + \varepsilon$. Dan met (45.9)

$$x_n = 1 - \frac{\varepsilon}{4} + \frac{\varepsilon}{4} 5^n.$$

Een zeer klein foutje in de startwaarden geeft dus al zeer snel een grote fout in het resultaat. Hetzelfde geldt voor afrondfoutjes bij het berekenen van $x_2, x_3, \ldots$ Een dergelijke methode zal men instabiel noemen.

Een minimale eis van stabiliteit is dat een foutje in een startwaarde of een afrondfoutje onderweg begaan, aanleiding geeft tot een begrensde fout verderop, zelfs als $h \to 0$, als er dus zeer veel recursie stappen tussen kunnen liggen. Hierbij is aangenomen dat voor $h \to 0$ ook niet de verstoring voldoende snel naar nul gaat, wat voor afrondfouten bv. een redelijke veronderstelling is.

Dit voorbeeld (denk aan (45.19)) motiveert nu:

(48.13) De multistep (47.11) heet stabiel indien de wortels van het bijbehorende polynoom ρ binnen of op de eenheidscirkel liggen en bovendien de wortels op de eenheidscirkel enkelvoudig zijn.

Tenslotte nog het begrip convergentie. Zoals we al in (48.12) aan een eenvoudig geval hebben gezien kan een multistep al dan niet convergeren afhankelijk van de gekozen startprocedure. Om moeilijkheden hiermee te vermijden definiëren we:

(48.14) <u>Definitie</u>. De multistep (47.11) heet <u>convergent</u> indien voor elk beginwaarde probleem waarvoor (46.8) geldt en iedere startprocedure, die voldoet aan

$$\lim_{h\to 0} x_i = x(t_0) \qquad i = 1,2,\ldots,k-1$$

de oplossing $\{x_n\}$ voldoet aan

$$\lim_{\substack{h\to 0\\ t_0+nh=t\in I}} x_n = x^*(t)$$

(48.15) <u>Stelling</u>. Voor convergentie is <u>nodig</u> stabiliteit en consistentie.

<u>Bewijs</u>.

a) We bewijzen eerst convergentie $\Rightarrow$ stabiliteit. Zij de multistep niet stabiel. Beschouw de diff.vgl. $x' = 0$, $x_0^* = 0$, zodat $x^*(t) = 0$. Er geldt nu $x_n+\alpha_1 x_{n-1}+\ldots+\alpha_k x_{n-k} = 0$. Nu heeft de recurrente betrekking $u_n+\alpha_1 u_{n-1}+\ldots+\alpha_k u_{n-k} = 0$ zeker hetzij een oplossing waarvoor $|u_n| = n$ (nl. als er een meervoudige wortel op de eenheidscirkel ligt) hetzij een oplossing $|u_n| = \mu^n$ met $\mu > 1$. In beide gevallen is ook $x_n = u_n\sqrt{h}$ een oplossing van de recursie en er geldt $x_0,\ldots,x_{k-1} \to 0$ voor $h \to 0$ en $x_n \to \infty$ voor $h \to 0$ en $nh \to t$. Derhalve geen convergentie.

b) Bewijs dat $\rho(1)=0$: Beschouw $x'=0$, $x_0^*=1$, zodat $x^*(t)=1$. De recursie luidt nu $x_n+\alpha_1 x_{n-1}+\ldots+\alpha_k x_{n-k}=0$. We nemen $x_0=x_1=\ldots=x_{k-1}=1$ voor elke h. Dit is een toegelaten startprocedure. De rij $\{x_n\}$ hangt nu niet van h af. Kies een $t\in I$ en laat h de rij t/n, $n=1,2,3,\ldots$ doorlopen. Dan moet x_n naar $x^*(t)=1$. Invullen in de recursie geeft $1+\alpha_1+\ldots+\alpha_k=0$.

c) Bewijs dat $\rho'(1) = \sigma(1)$, met de wetenschap dat reeds $\rho(1) = 0$ (zie b) en $\rho'(1) \neq 0$ (zie a). Beschouw $x' = 1$, $x_0 = 0$ zodat $x^*(t) = t$. Er geldt nu $x_n + \alpha_1 x_{n-1} + \ldots + \alpha_k x_{n-k} = h\sigma(1)$. ~~Beschouw nu de~~ Zij $x_n = n h \sigma(1)/\rho'(1)$ voor elke h. De waarden $x_0, \ldots, x_{k-1}$ voldoen aan de eis voor startwaarden. Voorts geldt $x_n + \alpha_1 x_{n-1} + \ldots + \alpha_k x_{n-k} =$

$$h \frac{\sigma(1)}{\rho'(1)} \Big[(n-k) \underbrace{(1 + \alpha_1 + \ldots + \alpha_k)}_{= \rho(1) = 0} + \underbrace{k + \alpha_1(k-1) + \ldots + \alpha_{k-1}}_{= \rho'(1)} \Big] = h\sigma(1)$$

zodat aan de recursie is voldaan. Kies weer t en $h = t/n$, $n = 1, 2, 3, \ldots$ Dan geldt $x_n = t\sigma(1)/\rho'(1)$ en dit moet naar t voor $n \to \infty$. Dus $\sigma(1) = \rho'(1)$.

§49. Convergentie.

In deze paragraaf zullen we de fout $x_n - x^*_n$ nader bezien. We zullen voldoende voorwaarden geven, waaronder de fout naar nul convergeert voor $h \to 0$.

We veronderstellen dat de verkregen numerieke oplossing (zo deze bestaat, wat we nog moeten bewijzen) voldoet aan:

(49.1) $$x_n + \alpha_1 x_{n-1} + \ldots + \alpha_k x_{n-k} = h[\beta_0 f(t_n, x_n) + \ldots + \beta_k f(t_{n-k}, x_{n-k})] + \varepsilon_n$$

waarbij in ε_n de afrondfouten kunnen worden opgenomen en de fouten die het gevolg zijn van het niet exact oplossen van de vergelijking bedoeld in (47.9).

De startwaarden $x_0, x_1, \ldots, x_{k-1}$ moeten t.o.v. $x^*_0, x^*_1, \ldots, x^*_{k-1}$ lichtelijk verstoord zijn, dus $x_i = x^*_i + e_i$ $(i = 0,1,\ldots,k-1)$.

Daarnaast weten we dat de exacte oplossing voldoet aan

$$x^*_n + \alpha_1 x^*_{n-1} + \ldots + \alpha_k x^*_{n-k} = h[\beta_0 f(t_n, x^*_n) + \ldots + \beta_k f(t_{n-k}, x^*_{n-k})] + h\delta_n.$$

Zij $e_n = x_n - x^*_n$. Dan voldoet $\{e_n\}$ aan de recursie

(49.2) $$e_n + \alpha_1 e_{n-1} + \ldots + \alpha_k e_{n-k} = h[\beta_0 f(t_n, x^*_n + e_n) - \beta_0 f(t_n, x^*_n) + \ldots\ldots + \beta_k f(t_{n-k}, x^*_{n-k} + e_{n-k}) - \beta_k f(t_{n-k}, x^*_{n-k})] + \varepsilon_n - h\delta_n$$

met de startwaarden $e_0, e_1, \ldots, e_{k-1}$.

We beschouwen het eenvoudige geval $x \in \mathbb{R}$.

(49.3) Stelling. Zij de methode consistent.

(i) Zij $A = \begin{pmatrix} 0 & 1 & & \Theta \\ & & \ddots & \\ \Theta & & & 1 \\ -\alpha_k & \ldots & \ldots & -\alpha_1 \end{pmatrix}$

(ii) Zij $\|A^n\| \leq M$ onafhankelijk van n, $M \geq 1$.

(iii) Zij $h \leq \dfrac{1}{2\beta_0 L}$ (L is de constante uit (46.7)).

(iv) Laat een rij $\varepsilon_k, \ldots, \varepsilon_n$ gegeven zijn zodat voor alle $n \leq N$ geldt dat $(t_n, x^*_n + e) \in \Omega$ voor alle e met

(v) $|e| \leq 4\, cMe^{M\gamma T}(\max\limits_{0 \leq i \leq k-1} |e_i| + \sum\limits_{i=k}^{n} |\varepsilon_i| + T\delta_{max})$ waarin

(vi) $\gamma = 2L \sum_{i=1}^{k} (|\beta_o \alpha_i| + |\beta_i|)$

(vii) $c = \sum_{i=1}^{k} (|\alpha_i| + h|\beta_i|L)$

(viii) $\delta_{max} = \max_{k \leq i \leq N} |\delta_i|$

Dan bestaat er een rij $x_o, \ldots, x_N$ die voldoet aan (49.1)⌈.
Voorts, als $x_o, \ldots, x_{i-1}$ voldoen aan (49.1) voor $n=k,\ldots,i-1$⌈ dan is (47.11) voor $n=i$ oplosbaar naar x_n en wel met de methode van successieve substitutie, die zeker convergeert binnen de grootste "bol" B (dit is in ons ééndimensionale geval een segment) met x_n^* als middelpunt, zodat $(t_n,u) \in \Omega$ voor alle $u \in B$.

Voor de genoemde rij $\{x_n\}$ geldt

(ix) $$|x_n - x_n^*| \leq 2Me^{M\gamma T}(\max_{0 \leq i \leq k-1} |e_i| + \sum_{i=k}^{n} |\varepsilon_i| + T\,\delta_{max}),$$

$k \leq n \leq N.$

Bewijs.

We nemen voorlopig de existentie van de rij $\{x_i\}$ aan. Dan volgt uit (49.2), als we nog schrijven $f(t_n, x_n^* + e_n) - f(t_n, x_n^*) = l_n e_n$, waarin l_n wel van e_n en x_n^* afhangt, maar voldoet aan $|l_n| \leq L$,

(49.4) $$(1-h\beta_o l_n)e_n + \ldots + (\alpha_k - h\beta_k\, l_{n-k})e_{n-k} - \varepsilon_n + h\delta_n = 0$$

Zij nu

(49.5) $$z_n = \begin{bmatrix} e_{n-k} \\ \vdots \\ e_{n-1} \end{bmatrix}, \quad b_n = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{\varepsilon_n - h\delta_n}{1-h\beta_o l_n} \end{bmatrix},$$

(49.6) $$A_n = \begin{bmatrix} 0 & 1 & & \Theta \\ & & \ddots & \\ & \Theta & & 1 \\ \dfrac{-\alpha_k + h\beta_k l_{n-k}}{1-h\beta_o l_n} & \cdots & & \dfrac{-\alpha_1 + h\beta_1 l_{n-1}}{1-h\beta_o l_n} \end{bmatrix}$$ (vgl. (45.14)).

⌈ met $x_i = x_i^* + e_i$ voor $i = o, \ldots, k-1$

Dan geldt

(49.7) $$z_{n+1} = A_n z_n + b_n$$
$$= b_n + A_n b_{n-1} + \dots + A_n A_{n-1} \dots A_k z_k.$$

(z_k is de vector der fouten in de beginwaarden) zodat

(49.8) $$\| z_{n+1} \| \leq \max_{k \leq p \leq n} \| A_n \dots A_p \| \; (\| z_k \| + \sum_{i=k}^{n} \| b_i \|).$$

Schrijven we nu $A_n = A + C_n$ dan geldt met de in (vi) genoemde γ :

(49.9) $$\| C_n \|_\infty \leq h\gamma .$$

Derhalve geldt, wegens (45.24)

(49.10) $$\max_{k \leq p \leq n} \| A_n \dots A_p \|_\infty \leq (1+Mh\gamma)^n M < M e^{M\gamma hn}$$

Dus

(49.11) $$\| z_{n+1} \|_\infty < M e^{M\gamma hn} (\| z_k \|_\infty + \sum_{i=k}^{n} \| b_i \|_\infty).$$

Nu geldt $\| b_i \| = \left| \dfrac{\varepsilon_i - h\delta_i}{1-h\beta_o l_i} \right|$ waarvan de noemer wegens (ii) minstens $\frac{1}{2}$ is zodat

(49.12) $$|e_n| \leq \| z_{n+1} \|_\infty \leq 2 M e^{M\gamma hn} (\| z_k \|_\infty + \sum_{k}^{n} | \varepsilon_i | + h \sum_{k}^{n} |\delta_i|)$$
$$\leq 2 M e^{M\gamma hn} (\| z_k \|_\infty + \sum_{k}^{n} | \varepsilon_i | + hn\, \delta_{max})$$

Wegens $hn \leq hN = T$ volgt hieruit (ix)

Blijft nog te bewijzen onze in het begin gemaakte aanname dat de rij $\{x_i\}$ bestaat en dat $(t_n, x_n) \in \Omega$ $(n=0,1,2,\dots N)$.

Laten (t_o, x_o), $(t_1, x_1) \dots$, $(t_{n-1}, x_{n-1}) \in \Omega$.

Dan is (ix) al vast waar met m i.p.v. n, $m < n$, en mh i.p.v T

We schrijven (49.2) als $e_n = \phi(e_n)$.

De "bol" B voor x_n gaat nu over in een bol $\tilde{B}$ voor e_n met 0 als middelpunt en als straal minstens het rechterlid van (v). ϕ is dus gedefinieerd op de bol $\tilde{B}$.

We gebruiken nu de volgende vaste puntstelling, een omformulering van NAI(26.6) voor het geval $k=\frac{1}{2}$.

(49.13) Zij ϕ gedefinieerd op een bol om 0 met straal r.

Zij $|\phi(x) - \phi(y)| \leq \frac{1}{2} |x-y|$ voor alle $x,y \in \tilde{B}$, en zij $|\phi(0)| \leq \frac{1}{2} r$.

Dan heeft de vergelijking $\phi(x) = x$ een oplossing binnen $\tilde{B}$, en de methode van successieve substitutie convergeert voor elk startpunt binnen $\tilde{B}$ naar deze wortel.

Aan de eis $|\phi(x)-\phi(y)| \leq \frac{1}{2}|x-y|$ is in ons geval zeker voldaan wegens (iii).

Voorts is

(49.14) $$\phi(0) = (-\alpha_1+h\beta_1 l_{n-1})e_{n-1}+\ldots+(-\alpha_k+h\beta_k l_{n-1})e_{n-k}+\varepsilon_n-h\delta_n$$

en dus

(49.15) $$|\phi(0)| \leq c.\max_{i<n}|e_i|+\varepsilon_n-h\delta_n \quad \text{(c als in (vii))}$$

$$\leq c.2Me^{M\gamma T}(\max_{0\leq i\leq k-1}|e_i|+\sum_{i=k}^{n-1}|\varepsilon_i|+(n-1)h\delta_{max})+\varepsilon_n-h\delta_n$$

dit mag wegens $c \geq 1$ (consistentie) →

$$\leq c.2Me^{M\gamma T}(\max_{0\leq i\leq k-1}|e_i|+\sum_{i=k}^{n}|\varepsilon_i|+nh\delta max)$$

$\leq$ halve straal van $\tilde{B}$.

Hiermee is alles bewezen.

(49.16) Het algemeen geval $x \in \mathbb{R}^m$ kan geheel analoog behandeld worden. We doen dit hier niet.

Opmerkingen

(49.17) Een multistep die aan de eisen van stelling (49.3) voldoet is blijkbaar convergent (zie def. (48.14)).

Aan de eisen van de stelling is voor voldoend kleine waarden van h zeker voldaan als a) $\|A^n\| \leq M$ onafhankelijk van n, b) er zijn geen afrondfouten, en c) $\delta_{max} \to 0$ voor $h \to 0$, maar dat zijn juist de eisen van stabiliteit en consistentie. Derhalve , samen met (48.15) de belangrijke eigenschap:

(49.18) Stelling. Stabiliteit en consistentie zijn noodzakelijk en voldoende voor convergentie van een multistep.

(49.19) Ω behoeft niet groter te zijn dan het gebied beschreven in (iv), zodat in feite slechts eisen aan f gesteld worden op een onmiddelijke omgeving van x^*.

(49.20) Voor begrensde T is het effect van een enkele fout ε_i begrensd, hoe groot het aantal stappen (d.w.z. hoe klein h) ook is. Hetzelfde is waar voor de fouten e_i in de startwaarden.

(49.21) Neem aan dat voor $h \to 0$ de discretisatiefout $O(h^p)$ is. Dan geldt m.b.v. van stelling (49.3) in afwezigheid van afrondfouten (dus alle $\varepsilon_i=0$) dat de approximatiefout $O(h^p)$ is.

Men interpretere dit niet verkeerd. Zij $x_h(t_n)$ resp. $x_{h/2}(t_{2n})$ een benadering voor $x^*(t)$ met de stapgrootten h en $h/2$. Stelling (49.3) zegt nu dat de afschatting voor $|x^*(t)-x_{h/2}(t_{2n})|$ een factor 2^p kleiner is dan de afschatting voor $|x^*(t)-x_h(t_n)|$ (onder de aanname $e_i = O(h^p)$ en $\varepsilon_i = 0$). Dit betekent niet:

$$|x^*(t)-x_{h/2}(t_{2n})| \sim 2^{-p} \cdot |x^*(t)-x_h(t_n)|.$$

Dat in de praktijk dit laatste toch redelijk blijkt te kloppen kan wel aannemelijk gemaakt worden. Men moet dan bewijzen dat

$$x_{t/h} - x^*(t) = h^p\, e(t) + O(h^{p+1}).$$

waarin e(t) niet van h afhangt. Voor eenstapsmethoden kan dit in het algemeen wel aangetoond worden. Voor multisteppen ligt het veel moeilijker. We bezien dit in §50.

De nu klassieke foutschattingen en convergentiebewijzen zijn voor het eerst gegeven door Dahlquist (1956, zie ook zijn dissertatie (1959)). De theorie van Dahlquist wordt zeer breedvoerig uiteengezet in de beide boeken van Henrici (zie literatuurlijst). Ansorge - Hass breidt de theorie verder uit naar partiële en speciaal hyperbolische differentiaalvergelijkingen.

§ 50 Nadere Foutbeschouwingen, asymptotisch gedrag van de fout.

(50.1) Doel van deze paragraaf is een beter inzicht te verkrijgen in het mechanisme van de foutvoortplanting bij multistepmethoden. In het bijzonder zal in vele gevallen een verscherping van de foutschattingen in (49.3) mogelijk zijn.
Dat dit wenselijk is, moge blijken uit de

(50.2) Opgave. De differentiaalvergelijking $\frac{dx}{dt} = -x$ wordt opgelost m.b.v. de multistep (47.8). Neem aan: $x(0) = 1$, $e_0 = e_1 = 0$. $\varepsilon_i = 0$ alle i. Laat zien dat (49.3) een relatieve fout van hoogstens 1% in x(12) garandeert indien $h < 10^{-6}$.

(50.3) In de praktijk zal voor het probleem uit (50.2) een stapgrootte $h \sim 10^{-2}$ al voldoen. De schatting (49.3) eist dus dat de hoeveelheid werk ongeveer 10.000 keer zo groot is als in werkelijkheid nodig is voor het geval uit (50.2). Dit alleen al motiveert zeker een nadere foutbeschouwing. We zullen daarom nu trachten niet een bovengrens (zoals in (49.3)) maar een echte schatting van de fout te geven voor voldoend kleine waarden van h.

(50.4). Zij de multistep van orde $p \geq 1$, dus $\delta_n = h^p x^{*(p+1)}(t_n) + O(h^{p+1})$.
We nemen aan dat ε_n (zie (49.1)) klein wordt t.o.v. $h\delta_n$ als $h \to 0$: $\varepsilon_n = O(h^{p+2})$. We nemen ook aan dat de startfouten e_i voldoen aan $e_i = O(h^{p+1})$, $i \leq k-1$.

Dan leert (49.3) (ix) dat $e_n = O(h^p)$. Dus mogen we in (49.2) $f(t_i, x_i^* + e_i) - f(t_i, x_i^*) = \left(\frac{\partial f}{\partial x}\Big|_{t,\xi}\right) e_i$ met $\xi \in (x_i^*, x_i^* + e_i)$ vervangen door $\left(\frac{\partial f}{\partial x}\Big|_{t,x_i^*}\right) e_i + O(h^p) e_i$. We definiëren $g(t) = \frac{\partial f}{\partial x}\Big|_{t,x^*(t)}$

Er geldt dus

(50.5) $$\sum_{i=0}^{k} \left[\alpha_i e_{n-i} - h\beta_i g_i e_{n-i} \right] + h^{p+1} C_p x_n^{*(p+1)} + O(h^{p+2}) = 0$$

waarbij in de term $O(h^{p+2})$ zijn opgenomen de ε_n, de O-term uit δ_n en de term $h\, O(h^p) e_i = O(h^{2p+1}) = O(h^{p+2})$.

Wegens $e_i = O(h^p)$ voeren we in $\tilde{e}_i = e_i / h^p$, dat is dus e_i bekeken door een vergrootglas. Als we ook nog bedenken dat $x_n^{*(p+1)} = x_{n-i}^{*(p+1)} + O(h)$ voor $i \leq k$ en dat $\sum_{i=0}^{k} \beta_i = \sigma(1)$, dan krijgen we uit (50.5):

(50.6) $$\sum_{i=0}^{k} \left[\alpha_i \tilde{e}_{n-i} - h\beta_i \left(g_i \tilde{e}_{n-i} - \frac{C_p}{\sigma(1)} x_{n-i}^{*(p+1)} \right) \right] + O(h^2) = 0.$$

Deze relatie doet sterk denken aan het oplossen met de steeds beschouwde multistep van de vergelijking

(50.7) $$\tilde{e}'(t) = g(t)\,\tilde{e}(t) - \frac{C_p}{\sigma(1)} x^{*(p+1)}(t).$$

Inderdaad valt (50.6) met (49.1) samen als we $\varepsilon_n = O(h^2)$ nemen (N.B. we bedoelen dus nu met (49.1) de bij (50.7) passende).

<u>We passen nu hiervoor (49.3) toe</u>! We krijgen dan uit (49.3)(ix):

(50.8) $$|\tilde{e}_n - \tilde{e}_n^*| \leq 2Me^{M\gamma T}\left(\max_{i\leq k-1} |\tilde{e}_i - \tilde{e}_i^*| + \sum_{i=k}^{n} O(h^2) + T h^p \tilde{e}^{*(p+1)}_{max}\right)$$

Beschouwen we (50.7) met beginwaarde $\tilde{e}_0^* = 0$, dan geldt $\tilde{e}_i^* = O(h)$ voor $i \leq k-1$, terwijl krachtens de aanname in het begin $\tilde{e}_i = O(h)$ voor $i \leq k-1$. Derhalve:

(50.9) $$\tilde{e}_n = \tilde{e}_n^* + O(h).$$

Voor de oorspronkelijke differentiaalvergelijking hebben we dus:

(50.10) <u>Stelling</u>. $x_n - x_n^* = h^p \tilde{e}^*(t_n) + O(h^{p+1})$, als $\tilde{e}^*(t_0) = 0$. ☒

Men noemt $\tilde{e}^*$ wel de <u>principal error function</u> en $C_p/\sigma(1)$ de fout constante (error constant) van de multistep.

(50.11) <u>Opgave</u> Reken, onder de aanname dat de $O(h^{p+1})$ term in (50.10) verwaarloosd mag worden, (50.3) na. (Het is niet eenvoudig a priori na te gaan of de bedoelde term inderdaad klein is; in de praktijk verschaft men zich hieromtrent vertrouwen door te kijken of bij gebruik van waarden h_1, h_2, h_3 van h geldt $[x_{t/h_1}(h_1) - x_{t/h_2}(h_2)]/[x_{t/h_2}(h_2) - x_{t/h_3}(h_3)] \approx (h_1^p - h_2^p)/(h_2^p - h_3^p)$;

hierin stelt $x(h_1)$ de met stapgrootte h_1 verkregen benaderde oplossing voor; idem $x(h_2)$ en $x(h_3)$.

(50.12) <u>Opmerking</u>. In (50.8) komt $\tilde{e}^{*(p+1)}$ voor, en het bestaan daarvan kunnen we wegens (50.7) alleen maar garanderen als x^* een $(2p+2)$-de afgeleide heeft. We kunnen echter wel met minder toe. Het is voldoende als x^* in $C^{(p+2)}$ zit: weliswaar mogen we dan in (50.8) de term $Th^p \tilde{e}^{*(p+1)}_{max}$ niet schrijven, maar wel mogen we dan schrijven $Th\, \tilde{e}^{*''}_{max}$, zoals men ziet analoog aan (48.7).

(50.13) <u>Opmerking</u>. Uit (50.7) vindt men (met de zgn. methode van variatie van constanten)

$$(50.14) \qquad \tilde{e}^*(t) = -\frac{C_p}{\sigma(1)} \int_{t_0}^{t} e^{\int_\tau^t g(s)ds}\, x^{*(p+1)}(\tau)\, d\tau$$

(hetgeen men door invullen gemakkelijk verifieert). Wanneer de differentiaalvergelijking $x' = f(t,x)$ nu eens lineair is: $x' = l(t)x + q(t)$, dan is $g(s) = l(s)$, en stelt de functie $t \mapsto e^{\int_\tau^t l(s)ds}$ de oplossing van de homogene differentiaalvergelijking $x' = l(t)x$ voor die in het punt τ de waarde 1 heeft. Door nu in (50.14) de $\int_{t_0}^t$ te vervangen door een Riemann som met maaswijdte h ziet men dat de fout $h\delta_i$ ($\approx C_p x^{*(p+1)}(t_i)$), gedeeld door $\sigma(1)$, zich volgens oplossingen van de homogene differentiaalvergelijking voortplant, juist zoals we dat in vroegere colleges voor speciale gevallen hebben aangetoond. Als deze oplossingen dalen **worden** eenmaal gemaakte fouten dus verzwakt voortgeplant, bij voldoend kleine h. Dit verklaart de discrepantie tussen (50.2) en (50.3).

§ 54. Beginwaarde problemen.

Eén van de meest eenvoudige beginwaardeproblemen, geassocieerd met een patiële differentiaalvergelijking is wel:

(54.1)
$$\begin{cases} u_t = u_{xx} \\ u(x,o) = f(x) \end{cases}$$

Gevraagd een oplossing van (54.1) voor $t \geqslant o$.
De differentiaalvergelijking uit (54.1) staat bekend als de warmte- of diffusievergelijking en is van het parabolische type.

(54.2) We veronderstellen f periodiek, periode 1 en $f \in C^2$.
De Fourier-reeks van f convergeert dan zeker uniform.

(54.3) Het ligt voor de hand (54.1) m.b.v. differenties te benaderen op een rooster, maaswijdte Δt in de t-richting, Δx in de x-richting. Dan is

(54.4)
$$\frac{U_j^{n+1}-U_j^n}{\Delta t} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{(\Delta x)^2}$$

een benadering voor de differentiaalvergelijking uit (54.1). Hierin is U_j^n de waarde van de roosterfunctie in $(j\Delta x, n\Delta t)$. (54.4) noemt men wel een differentieschema. Een ander zeer bekend schema is dat van Crank en Nicholson:

(54.5)
$$\frac{U_j^{n+1}-U_j^n}{\Delta t} = \tfrac{1}{2}\left\{\frac{U_{j+1}^{n+1}-2U_j^{n+1}+U_{j-1}^{n+1}}{(\Delta x)^2} + \frac{U_{j+1}^n-2U_j^n+U_{j-1}^n}{(\Delta x)^2}\right.$$

(54.6) Op de gebruikelijke wijze wordt consistentie gedefinieerd. Zo b.v. voor (54.4): Men substituere $u(j\Delta x,n\Delta t)$ voor U_j^n met u de oplossing van (54.1).

(54.7). <u>Definitie</u> De discretisatiefout van (54.4) in $(j\Delta x, n\Delta t)$ is:

$$\delta_j^n = \frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2}$$

(54.4) heet consistent indien $\lim_{\substack{\Delta x \to o \\ \Delta t \to o}} \delta_j^n = o$, $n\Delta t$ en $j\Delta x$ vast.

Nu is het duidelijk dat voor $u \in C^4$ geldt

$$\delta_j^n = O(\Delta t) + O((\Delta x)^2).$$

Geheel analoog wordt de discretisatiefout voor het Crank-Nicholson-schema gedefinieerd. Evenzo consis...

<u>Opgave.</u>
(54.8) Ga na dat de discr.fout van Crank-Nicholson

$$O((\Delta t)^2) + O((\Delta x)^2) \text{ is.}$$

Aanw: ontwikkel in Taylor rond ($j\Delta x$, $(n+\frac{1}{2})\Delta t$).

(54.9) Zoals altijd is het probleem convergentie aan te ton... d.w.z. te bewijzen

$$\lim_{\substack{\Delta x \to o \\ \Delta t \to o}} U_j^n = u(x,t) \quad n\Delta t = t \;,\; j\Delta x = x$$

(waarbij i.h.a. tussen Δx en Δt een zeker verband zal bestaan).

Evenals bij gewone eerste orde differentiaalvergelij... speelt het begrip stabiliteit een beslissende rol. Nemen we voor het vervolg aan $\Delta x = \frac{1}{N}$, N natuurlijk ... Door te eisen dat de roosterfunctie eveneens periodiek is met periode 1, dus $U_{j+N}^n = U_j^n$ voor alle j, kunnen we alle differentievergelijkingen (54.4) samenvatten al...

(54.10) $$U^{n+1} = C.U^n$$

waarin U^n de vector met N componenten is, met als j... coördinaat U_{j-1}^n. C is een matrix, die we nu niet ve... zullen bezien.

(54.11) Opgave Ga na dat het Crank-Nicholson schema te schrijven is als

$$A\ U^{n+1} = B\ U^n$$

(en dus ook weer als $U^{n+1} = C\ U^n$, indien A nonsingulier met een van (54.10) verschillende matrix C).

(54.12) Definitie (54.4) heet stabiel indien $\|C^n\|$ (zie (54.10)) uniform begrensd is voor alle n en Δt met $o \leq n\Delta t \leq T$, $T > o$.

Algemeen heet een schema stabiel indien het op de gedaante (54.10) gebracht kan worden met $\|C^n\|$ uniform begrensd zoals in (54.12).

(54.13) We onderzoeken eerst de stabiliteit; daarna zullen we aantonen dat voor die combinaties van Δt en Δx waarvoor (54.4) stabiel is ook convergentie optreedt (en even voor Crank-Nicholson).

We maken nu gebruik van de periodiciteit.
Zoals bekend kan de exacte oplossing geschreven worden als een Fourier-reeks. Dit doen we ook voor de opl. op het rooster. We stellen

(54.14)
$$U_j^n = \sum_{k=o}^{N-1} \hat{u}^n(k)\ e^{ik\pi j/N}.$$

Er geldt:

(54.15)
$$\sum_{j=o}^{N-1} e^{ik\pi j/N} \cdot e^{-im\pi j/N} = \begin{cases} o & k \neq m \\ N & k = m. \end{cases}$$

$k,m = o,1,2,\ldots,N-1$
waaruit volgt:

(54.16)
$$\hat{u}^o(k) = \frac{1}{N} \sum_{j=o}^{N-1} U_j^0\ e^{-ik\pi j/N}.$$

We kunnen dit ook als volgt interpreteren: definieer in de lineaire ruimte van de functies gedefinieerd op de ... j/N, $j=o,1,\ldots,N-1$ het unitaire inprodukt:

(54.17) $(f,g) = \frac{1}{N} \sum_{j=o}^{N-1} f(j/N)\, \overline{g(j/N)}$.

Dan is $\hat{u}^0(k) = (U^0, e^{ik\pi x})$ terwijl (54.15) aantoont dat $1, e^{i\pi x}, e^{2i\pi x}, \ldots, e^{(N-1)i\pi x}$ een orthogonale basis t.a.v. dit inprodukt vormen. Dus

$$U_j^0 = \sum_k (U^0, e^{ik\pi x})\, e^{ik\pi j/N}$$

en $(U^0,U^0) = \|U^0\|^2 = \sum_k (U^0, e^{ik\pi x})^2 . N = N \sum_k (\hat{u}^0(k))^2$.

Deze laatste relatie staat bekend onder de naam Parceval-relatie.

(54.18) Substitueren we (54.14) in (54.4) dan vinden we voor $\hat{u}^n(k)$ de recursie:

(54.19)
$$\hat{u}^{n+1}(k) = G(k,\Delta t)\, \hat{u}^n(k)$$

met als <u>amplificatiefactor</u> G :

(54.20)
$$G(k,\Delta t) = 1 - 4 . \frac{\Delta t}{(\Delta x)^2} \sin^2 \frac{k\pi}{2N}.$$

(54.21) <u>Stelling</u> (54.10) is stabiel indien $|G^n(k,\Delta t)|$ uniform begrensd is voor alle n en Δt, $o < n\Delta t < T$ en alle k. $\|U^n\| = \|U^n\|_2$.

<u>Bewijs:</u> $U_j^n = (C^n U^0)_j = \sum_k G^n(k,\Delta t)\, \hat{u}^0(k)\, e^{ik\pi j/N}$

$$(U_j^n, U_j^n) = N . \sum_k [G^n(k,\Delta t)\, \hat{u}^0(k)]^2$$

$$\leq M.N. \sum_k |\hat{u}^0(k)|^2 \leq M^2 (U^0, U^0).$$

indien M de uniforme bovengrens voor $|G^n(k,\Delta t)|$ is.

Dus:
$\frac{\|C^n U_0\|}{\|U_0\|} \leq M$ waaruit volgt $\|C^n\|_2 \leq M$.

(54.22) <u>Opgave</u> Ga na dat de voorwaarde van (54.21) ook nodig is.

Voor (54.20) betekent de voorwaarde van (54.21) kennelijk

$$\frac{\Delta t}{(\Delta x)^2} \leqslant \tfrac{1}{2}.$$

Onder deze voorwaarde (dus stabiliteit) volgt eenvoudig de eerder aangekondigde convergentie van (54.4).
Zij nml. e^n de vector (N coördinaten):

$$e^n = U^n - u^n$$

waarin de j^{de} coördinaat van u^n gegeven wordt door $u((j-1)\frac{1}{N}, n\Delta t)$. e^n stelt dus de fout per horizontale rij voor. Dan volgt uit (54,10) en (54.6):

$$e^{n+1} = C\, e^n + \Delta t.\delta^n$$

waarin δ^n de vector van de discr. fouten voorstelt.
Dus als $\|C^n\|_2 \leqslant M$:

$$e^{n+1} = C^{n+1}\, e^0 + \Delta t[\, C^n\delta^1 + C^{n-1}\delta^2 + \ldots + \delta^n]$$

$$\|e^{n+1}\|_2 \leqslant M\|e^0\|_2 + \Delta t.M.n.\ \max \|\delta^n\|_2$$

$$\leqslant M\|e^0\|_2 + M.T.\ \max \|\delta^n\|_2$$

waaruit de convergentie volgt.

(54.23) We hebben nu convergentie, maar niet in de sup-norm.
Maken we gebruik van de periodiciteit en $f \in C^{(2)}$ dan impliceert convergentie in de gebruikte norm ook uniforme convergentie, d.w.z. convergentie in de sup-norm (∞- norm).
Maar we kunnen niet zeggen dat $\|e^n\|_\infty = O(\max\limits_{0 \leqslant i \leqslant n} \|\delta^i\|_\infty)$.
Dan zouden we moeten weten of $\|C^n\|_\infty$ uniform in n begrensd is

(54.24) Op het Crank-Nicholson schema kunnen we onze theorie ook toepassen. Zonder eerst na te gaan of A (zie opgave (54.11)) non-singulier is, substitueren we weer (54.14) nu in (54.5).
Dan vinden we voor de $\hat{u}^n(k)$ de recursie:

(54.25)
$$\hat{u}^{n+1}(k) = \frac{1-2\frac{\Delta t}{(\Delta x)^2}\sin^2\frac{k\pi}{2N}}{1+2\frac{\Delta t}{(\Delta x)^2}\sin^2\frac{k\pi}{2N}}\hat{u}^n(k)$$

zodat de amplificatiefactor aan de eis van (54.21) voldoet voor alle $\frac{\Delta t}{(\Delta x)^2}$. Dit impliceert dat het Crank-Nicholson-schema een unieke oplossing bezit en dus dat A nonsingulier is voor alle $\frac{\Delta t}{(\Delta x)^2}$. en ook $\|(A^{-1}B)^n\|_2$ uniform in n begrensd is voor alle n. $(0 < n\Delta t < T)$.

(54.26) De beschreven methode is natuurlijk alleen toepasbaar bij een differentiaalvergelijking met constante coëfficiënten. Maar dan loont het ook wel de moeite. Bezie b.v. de warmtevergelijking in twee dimensies:

(54.27)
$$u_t = u_{xx} + u_{yy}$$

waarbij als startwaarde een functie f(x,y) gegeven is, periodiek in x- en y-richting met periode 1. Het is eenvoudig een differentieschema analoog aan (54.4) op te stellen en dit kan ook nog in de vorm (54.10) geschreven worden. Analoog Crank-Nicholson.
De optredende matrices C zijn echter bijzonder onplezierig terwijl de methode m.b.v. Fourierreeksen eenvoudig werkt:
Zij nml.

$$\frac{U^{n+1}_{j,k} - U^n_{j,k}}{\Delta t} = \frac{U^n_{j+1,k} - 2U^n_{j,k} + U^n_{j-1,k}}{(\Delta x)^2} + \frac{U^n_{j,k+1} - 2U^n_{j,k} + U^n_{j,k-1}}{(\Delta x)^2}$$

het analogon van (54.4). $U^n_{j,k}$ is de waarde van de roosterfunctie in $(j\Delta x, k\Delta x, n\Delta t)$, $\Delta x = \frac{1}{N}$.

Stel nu $U^n_{j,k} = \sum_{p,q} \hat{u}^n(p,q)\, e^{ip\pi j/N} . e^{iq\pi k/N}$

We vinden voor $\hat{u}^n(p,q)$ de recursie:

$$\hat{u}^{n+1}(p,q) = (1 - 4\frac{\Delta t}{(\Delta x)^2} \sin^2 \frac{p\pi}{2N} - 4\frac{\Delta t}{(\Delta x)^2}\sin^2 \frac{q\pi}{2N}\ )\ \hat{u}^n(p,q).$$

en geheel analoog aan het voorgaande is er stabiliteit en convergentie indien

$$|(1-4\frac{\Delta t}{(\Delta x)^2} \sin^2 \frac{p\pi}{2N} - 4\frac{\Delta t}{(\Delta x)^2} \sin^2 \frac{q\pi}{2N}\ )^n|$$

uniform in n begrensd voor alle p en q, dus $\frac{\Delta t}{(\Delta x)^2} \leqslant \frac{1}{4}$.

(Merk op dat de verhouding tussen Δt en Δx hier veel ongunstiger is dan voor het schema (54.4)).

(54.28) Men kan ook de methode toepassen op niet parabolische differentiaalvergelijkingen, b.v. de (hyperbolische) golfvergelijking (in één dimensie):

(54.29)
$$u_{tt} = u_{xx}.$$

met voor t=o gegeven u(x,o) en $u_t(x,o)$ die beide weer periodiek, periode 1 verondersteld worden.
We voeren weer roosterfuncties in op een orthogonaal rooster met maaswijdte Δx in de x-richting ($\Delta x = \frac{1}{N}$, N natuurlijk getal) en Δt in de t-richting. Als differentieschema kiezen we:

(54.30)
$$\frac{U^{n+1}_j - 2U^n_j + U^{n-1}_{j-1}}{(\Delta t)^2} = \frac{U^n_{j+1} - 2U^n_j + U^n_{j+1}}{(\Delta x)^2}$$

In tegenstelling tot de warmtevergelijking is de golfvergelijking een differentiaalvergelijking met een <u>tweede</u> orde afgeleide in de t-richting. Dit komt in de benaderende differentievergelijkingen (54.30) tot uiting hierin (naast het feit dat (54.30) een multistepmethode is) dat (54.30) niet begrensde oplossingen bezit voor $n\Delta t \in [o,T]$ met wel begrensde startwaarden, kies maar $U^0_j = o$ en $U^1_j = 1$ voor alle j en laat $\Delta t \to o$ gaan. We zullen dus geen stabiliteit in de zin van (54.10) mogen verwachten.

We omzeilen het probleem door (54.29) terug te brengen tot een stelsel eerste orde partiële differentiaalvergelijkingen en (54.30) tot een differentieschema voor dit stelsel.
Van dit equivalente schema zullen we stabiliteit en dus convergentie aantonen (en dus geen stabiliteit van (54.30)).
Tenslotte tonen we wel convergentie van (54.30) aan, uitgaande van de dan al bewezen convergentie voor het differentieschema dat de eerste orde differentiaalvergelijkingen benadert.
(54.29) is equivalent met

(54.31) $$\begin{cases} v_x = w_t \\ v_t = w_x \end{cases}$$

door te stellen $w = u_x$ en $v = u_t$

Het differentieschema (54.30) is equivalent met

(54.32a) $$\frac{V_j^{n+1} - V_j^n}{\Delta t} = \frac{W_{j+\frac{1}{2}}^n - W_{j-\frac{1}{2}}^n}{\Delta x}$$

(54.32 b) $$\frac{W_{j-\frac{1}{2}}^{n+1} - W_{j-\frac{1}{2}}^n}{\Delta t} = \frac{V_j^{n+1} - V_{j-1}^{n+1}}{\Delta x}$$

door te stellen

$$V_j^n = \frac{U_j^n - U_j^{n-1}}{\Delta t}.$$

$$W_{j-\frac{1}{2}}^n = \frac{U_j^n - U_{j-1}^n}{\Delta x}$$

Zij $\left(\frac{V^n}{W^n}\right)$de gepartitioneerde vector met als j^{de} coördinaat van V^n : V_{j-1}^n en als j^{de} coördinaat van W^n : $W_{j-\frac{1}{2}}^n$.
Dan is er een matrix C zo dat

(54.33) $$\left(\frac{V^{n+1}}{W^{n+1}}\right) = C . \left(\frac{V^n}{W^n}\right)$$

We tonen nu eerst aan dat onder zekere voorwaarden (54.32) een stabiel schema is.
We schrijven de oplossing van (54.32) als:

(54.34a) $$V_j^n = \sum_{k=0}^{N-1} \hat{v}^n(k)\, e^{ik\pi j/N}$$

(54.34b) $$W_{j+\frac{1}{2}}^n = \sum_{k=o}^{N-1} \hat{w}^n(k)\, e^{ik\pi(j+\frac{1}{2})/N}$$

geheel analoog aan (54.14). Nu is (54.34) oplossing indien voldaan is aan de recurrente betrekking.

(54.35) $$\begin{bmatrix} \hat{v}^{n+1}(k) \\ \hat{w}^{n+1}(k) \end{bmatrix} = G(k,\Delta t) \begin{bmatrix} \hat{v}^n(k) \\ \hat{w}^n(k) \end{bmatrix}$$

waarin G, de <u>amplificatiematrix</u> gegeven wordt door

(54.36) $$G = \begin{bmatrix} 1 & i\alpha \\ i\alpha & 1-\alpha^2 \end{bmatrix} \qquad \alpha = 2\,\frac{\Delta t}{\Delta x}\,\sin\frac{k\pi}{2N}\,.$$

(54.37) Stelling (54.21) is in iets gewijzigde vorm toepasbaar; het verschil is dat we nu vectoren i.p.v. scalairen hebben. Omdat het bewijs van (54.21) uitsluitend op de Parseval-relatie berust , bezien we deze relatie nog eens voor ons geval apart. Daartoe herschrijven we (54.34) als

$$V_j^n = \sum_{k=o}^{N-1} (V^n, e^{ik\pi x})\, e^{ik\pi j/N}$$

$$W_{j+\frac{1}{2}}^n = \sum_{k=o}^{N-1} \langle W^n, e^{ik\pi x}\rangle\, e^{ik\pi(j+\frac{1}{2})/N}$$

waarin (,) en $\langle\ ,\ \rangle$ inprodukten zijn,

(,) gedefinieerd als (54.17) en $\langle\ ,\ \rangle$ als

$$\langle f,g \rangle = \sum_{j=o}^{N-1} f\left(\frac{j+\frac{1}{2}}{N}\right) \overline{g\left(\frac{j+\frac{1}{2}}{N}\right)}.$$

Dan luidt de Parseval-relatie:

$$\|\left(\frac{V^n}{W^n}\right)\|_2^2 = \|V^n\|_2^2 + \|W^n\|_2^2 = N\left(\sum_{k=0}^{N-1} (|\hat{v}^n(k)|^2 + |\hat{w}^n(k)|^2)\right).$$

Het analogon van (54.21) blijft gelden met

$\|G^n(k,\Delta t)\|_2$ uniform begrensd (i.p.v. $|G^n(k,\Delta t)|$)

en met als matrix-norm van C de geassocieerde van de 2-norm (in de $\mathbb{R}^N \times \mathbb{R}^N$).

(54.38) Om de stabiliteit te onderzoeken passen we (45.18) op $G(k,\Delta t)$ toe. Nu worden de eigenwaarden van G gegeven door

$$\lambda_{1.2} = \frac{2-\alpha^2 \pm \sqrt{\alpha^4 - 4\alpha^2}}{2}$$

Voor α=o geldt $G \equiv I$ (2x2 eenheidsmatrix).

Dus is (54.32) stabiel voor $\alpha \in (-2,+2)$, d.i.

$\frac{\Delta t}{\Delta x} < 1$. Voor $\frac{\Delta t}{\Delta x} = 1$ is de Jordan-normaalvorm van G niet voor alle k een diagonaalmatrix, dus geen stabiliteit voor $\frac{\Delta t}{\Delta x} = 1$.

(54.39) Analoog aan wat voor de warmtevergelijking is afgeleid vinden we:

$$\left(\frac{\Delta t}{\Delta x} < 1\right)$$

$$\|\left(\frac{e_v^n}{e_w^n}\right)\|_2 \leq M.\ \|\left(\frac{e_v^0}{e_w^0}\right)\|_2 + M.T.\ \max_p \|\left(\frac{\delta_v^p}{\delta_w^p}\right)\|_2$$

waarin $\left(\frac{e_v^n}{e_w^n}\right)$ de vector van de fouten op de n-de horizontale rij voorstelt; $\left(\frac{\delta_v^n}{\delta_w^n}\right)$ die der discretisatiefouten.

(54.40) Dit toont nog eens de convergentie van het schema (54.32) aan, uitgaande van de stabiliteit van dit schema. Als de exacte oplossing (v,w) en zijn partiële afgeleide (v_x,w_x) continu en in de x-richting periodiek zijn volgt ook dat de benaderde oplossing naar de exacte convergeert in de sup-norm (dus uniform).

We weten echter niet of de fout als $O(\Delta t^\alpha)$, voor zekere $\alpha > o$, naar nul gaat.

We herinneren ons nog: $U_j^n = U_j^{n-1} + \Delta t.V_j^n$

wat een benaderde integratieformule voor u_t=v is.

Als nu maar V_j^n convergeert naar $v(j/N,n\Delta t)$ voor $\Delta t \to o$

z dan is wel duidelijk dat U_j^n naar $u(j/N, n\Delta t)$ convergeert.

(54.41) Een andere situatie waarin (54.29) optreedt is de volgende: gegeven $u(x,o)$ en $u_t(x,o)$ voor $x \in [o,L]$ terwijl een oplossing gevraagd wordt op de driehoek met hoekpunten (o,o) , (o,L) en $(\frac{1}{2}L,p)$, $|p| \leq \frac{1}{2}L$. Dit geval brengen we tot het vorige terug door $u(x,o)$ en $u_t(x,o)$ tot een periodieke functie uit te breiden, periode $> L$. Voor dit uitgebreide probleem is convergentie verzekerd.

Men kan echter eenvoudig aantonen dat op de beschouwde driehoek de oplossing onafhankelijk is van de uitbreidingen van $u(x,o)$ en $u_t(x,o)$.

Dus ook convergentie voor het probleem op de driehoek mits

$\frac{\Delta t}{\Delta x} < 1.$

(54.42) Indien $\frac{\Delta t}{\Delta x} = 1$ kan ook convergentie aangetoond worden.

Zij L=1. Er geldt:

$$e_j^n = e_{j+1}^{n-1} - e_j^{n-2} + (\Delta t)^2\ \delta_j^n$$

waarin e_j^n de fout in U_j^n voorstelt en δ_j^n de discretisatiefout.

Herhaald uitschrijven en afschatten:

$$|e^n_j| \leq |\sum_{p=o}^{n-1} e^{1}_{j-n+1+2p} - \sum_{p=1}^{n-2} e^0_{j-n+2p}| + \frac{n(n-1)}{2}\,\Delta t^2 \max|\delta_j^n|\ .$$

Nu is $e_j = o$ voor alle j (afgezien van afrondfouten)

en zij $e^1_j = O(\Delta t^\alpha)$ voor alle j. Dan:

$$|e_j^n| \leq n.\max|e_j^1| + \frac{n(n-1)}{2}(\Delta t)^2 \max|\delta^n|$$

$$\leq O(\Delta t^{\alpha-1}) + O(\max|\delta^n|).$$

Als nu $u \in C^{(4)}$ dan $O(\max|\delta_j^n|) = O((\Delta t)^2)$

zodat als $\alpha > 1$ convergentie verzekerd is.

(54.43) Opgave. We berekenen de startwaarden U_j^1 als volgt:
$\frac{U_j^1 - U_j^{-1}}{2\Delta t} = u_t(j/N,o) + O((\Delta t)^2)$. U_j^{-1} elimineren we m.b.v. (54,30). Ga na dat op deze wijze een U_j^1 verkregen wordt die $u(j/N,\Delta t)$ benadert met een fout $O((\Delta t)^3)$.$(\Delta x = \Delta t = 1/N)$.

(54.44) Sommige auteurs noemen het schema (54.30) stabiel voor $\frac{\Delta t}{\Delta x} \leqslant 1$. Deze hebben dan een andere stabiliteitsdefinitie (zie b.v. Forsythe en Wasow blz. 32).
Merk nogmaals op dàt we hier alleen stabiliteit voor een differentieschema gedefinieerd hebben dat een differentiaalvergelijking met eerste orde afgeleide in de t-richting benadert. De gegeven theorie m.b.v. Fourier-reeksen en de gegeven stabiliteitsdefinitie is afkomstig van Lax en Richtmyer. Zie vooral het boek van Richtmeyer en Morton.

Literatuur:

R. Ansorge, R.Hass — Convergenz von Differenzenverfahren. Springer Verlag, Lecture Notes (1970).

G.E. Forsythe, W.R. Wasow — Finite Difference Methods for Partial Differential Equations. John Wiley & Sons, New York (1960)

R.D. Richtmyer, K.W. Morton. — Difference Methods for Initial-Value Problems 2nd ed. Interscience. New York (1967).

Numerieke behandeling van differentiaalvergelijkingen

1. Probleemstelling.

We bekijken de eerste orde diff. verg :

(1.1) $$\frac{dx}{dt} = f(t, x) \quad , \quad x \in \mathbb{R}.$$

met $x(t_0) = x_0$ gegeven.

Zij L de Lipschitz-constante van f op het ons interesserende gebied. We nemen aan dat de oplossing van (1.1) bestaat en eenduidig is op $[t_0, t_0 + T]$.
We trachten een benaderde oplossing te vinden voor $t \in [t_0, t_0 + T]$. We kunnen echter nooit op een numerieke manier voor alle $t \in [t_0, t_0 + T]$ de oplossing vinden, maar zullen ons moeten beperken tot een benadering op zekere zelf te kiezen punten t_n We zullen t ook vaak als de tijd interpreteren, en dan spreken over de tijdstippen t_n.

2. Numerieke oplosmethoden

(2.1). We definiëren een stapgrootte $h = T/N$, N een natuurlijk getal. Met X_n^* resp X_n geven we de waarde aan van de exacte resp. van een benaderde oplossing in het punt $t_n = t_0 + n.h$. Een aantal oplosmethoden berusten op de opmerking dat

(2.2) $$X_n^* = X_{n-p}^* + \int_{t_{n-p}}^{t_n} f(t, x^*)\, dt.$$

Nu geldt voor een willekeurige functie $g \in C^1$ bijv.

$$(2.3) \qquad \int_{t_{n-1}}^{t_n} g(t)dt = hg(t_{n-1}) + \frac{h^2}{2} g'(\xi_n) \qquad t_{n-1} < \xi_n < t_n$$

(integratie van het 0e orde interpolatie polynoom in t_{n-1})

Door toepassing op (2.2) komt er

$$(2.4) \qquad x_n^* = x_{n-1}^* + h\, f_{n-1}^* + \frac{h^2}{2} x^{*''}(\xi_n), \qquad t_{n-1} < \xi_n < t_n$$

Analoog krijgt men met een kwadratuurformule als erachter aangeduid:

$$(2.5) \qquad x_n^* = x_{n-1}^* + h\, f_n^* - \frac{h^2}{2} x^{*''}(\xi_n) \qquad (0^e \text{ gr. interp. pol. in } t_n)$$

$$(2.6) \qquad x_n^* = x_{n-1}^* + \tfrac{1}{2}h[f_n^* + f_{n-1}^*] - \frac{h^3}{12} x^{*'''}(\xi_n) \quad \text{(trapezium)}$$

$$(2.7) \qquad x_n^* = x_{n-1}^* + \tfrac{1}{2}h[3f_{n-1}^* - f_{n-2}^*] + \frac{5}{12}h^3 x^{*'''}(\xi_n) \quad (1^e \text{ gr.interp.pol. in } t_{n-1}, t_{n-2})\ ^{*)}$$

$$(2.8) \qquad x_n^* = x_{n-2}^* + 2h\, f_{n-1}^* + \frac{h^3}{3} x^{*'''}(\xi_n) \quad \text{(midpoint)}$$

$$(2.9) \qquad x_n^* = x_{n-2}^* + \tfrac{1}{3}h[f_n^* + 4f_{n-1}^* + f_{n-2}^*] - \frac{h^5}{90} x^{*(5)}(\xi_n) \quad \text{(Simpson)}$$

Hierin is $f_n^* = f(t_n, x_n^*)$, etc. Door in (2.4) t/m (2.9) de sterretjes weg te laten, en ook de termen met ξ_n, worden dan blijkbaar benaderde oplossingen van de dvgl. gegenereerd, waarbij de waarden $x_{n-1}, x_{n-2}, \ldots$ een waarde x_n bepalen. Men noemt dergelijke betrekkingen wel recurrente betrekkingen.

Formules (2.4) en (2.5) duidt men wel aan als Eulèr resp. Euler backwards.

Het valt op dat:

- de methoden niet alle dezelfde orde van nauwkeurigheid hebben
- (2.5), (2.6) en (2.9) x_n niet zondermeer opleveren, omdat x_n ook in het rechterlid voorkomt. Deze formules heten daarom <u>impliciet</u>, in tegenstelling tot de overige, die <u>expliciet</u> heten.

---

*) Dit interpolatie polynoom wordt van t_{n-1} tot t_n geïntegreerd; dat is dus eigenlijk extrapolerend integreren.

Bij de impliciete methoden bepaalt men x_n meestal met successieve substitutie (zie ook An II (6.22)). Bijv. voor (2.6):

Zij $x_n^{(0)}$ een beginschatting voor x_n (bijv. $x_n^{(0)} = x_{n-1}$). Dan genereert men $x_n^{(1)}, x_n^{(2)}, \ldots$ d.m.v.

$x_n^{(i+1)} = x_{n-1} + \frac{1}{2}h(f(t_n, x_n^{(i)}) + f(t_{n-1}, x_{n-1}))$. Voor h klein genoeg convergeert dit: $|x_n^{(i+1)} - x_n^{(i)}| \leq \frac{1}{2}h\ L|x_n^{(i)} - x_n^{(i-1)}|$ (L de Lipschitz const) zodat voor $\frac{1}{2}h\ L \leq q < 1$ geldt $|x_n^{(i)} - x_n^{(j)}| \leq \frac{q^j}{1-q}\ |x_n^{(1)} - x_n^{(0)}|$, als $i > j$, dus Cauchy's criterium is vervuld. De convergentie is zelfs snel omdat beginschattingen als $x_n^{(0)} = x_{n-1}$ bij kleine stapgrootte uiteraard nogal goed zijn. De reden voor impliciete methoden is o.a. dat men op deze wijze interpolerend integreert, in tegenstelling tot expliciete methoden, waarbij men extrapolerend integreert, wat wel onnauwkeuriger zal zijn.

(2.10) De vraag of een hogere orde methode ook een nauwkeuriger antwoord oplevert, vereist nader onderzoek. (Zie hiervoor bijv. Num. An. I)

(2.11) De verkregen benaderingsmethoden hebben alle de vorm

(2.12) $$x_n + \alpha_1\ x_{n-1} + \ldots + \alpha_k\ x_{n-k} = h\ [\beta_0 f_n + \ldots + \beta_k f_{n-k}]$$

Zo'n algorithme noemt men een multistep of ook een k-staps methode. Voor k = 1 staat er een onestep.

(2.13) Zij $\delta_n = \frac{1}{h}\ \{x_n^* + \alpha_1 x_{n-1}^* + \ldots + \alpha_k x_{n-k}^* - h[\beta_0 f_n^* + \ldots + \beta_k\ f_{n-k}^*]\}$. Dan heet δ_n de locale discretisatiefout (ook wel truncation error). Zo is dus bij (2.6) de locale discretisatiefout $-\frac{h^2}{12}\ x^{*'''}(\xi)$.

De reden voor het delen door h bij de definitie van locale discretisatiefout is dat (2.12) eigenlijk t.o.v. de differentiaalvergelijking met h vermenigvuldigd is. Dit is bijv. heel duidelijk bij de regel van Euler: $x_n = x_{n-1} + h\ f$, oftewel $x_n - x_{n-1} = h\ f$,

terwijl een benadering van de dvgl. eigenlijk zou luiden $(x_n - x_{n-1})/h = f$. De locale discretisatiefout geeft dus in feite aan hoe goed (2.12) de dvgl. benadert.

(2.14) Tot dusverre werd het begrip orde in een intuitieve zin gebruikt. We geven nu een definitie.

Als x^* voldoende vaak differentieerbaar is heeft δ_n steeds de gedaante $Ch^p x^{*(p+1)} + O(h^{p+1})$. Men ziet dit in door in (2.13) in te vullen $x^*_{n-j} = x^*_n - \frac{jh}{1!} x^{*'}_n + \frac{(jh)^2}{2!} x^{*''}_n - ..$ en $f^*_{n-j} (= x^{*'}_{n-j}) = x^{*'}_n - \frac{jh}{1!} x^{*''}_n + \ldots$ Men zegt dan dat (2.12) van de orde p is, als p de grootste waarde is zodat δ_n de gegeven gedaante heeft voor alle voldoend vaak differentieerbare x.

(2.15) Het vervelende van multisteps is dat men er een startprocedure voor nodig heeft om $x_1, \ldots, x_{k-1}$ te bepalen, aangezien alleen x_0 gegeven is.

(2.16) Opgave: Methode (2.12) is dan en slechts dan van de orde $p \geq 1$ als geldt:

$$\begin{cases} 1 + \alpha_1 + \cdots + \alpha_k = 0 \\ k + (k-1)\alpha_1 + (k-2)\alpha_2 + \cdots + \alpha_{k-1} = \beta_0 + \cdots + \beta_k \end{cases}$$

Deze eigenschap noemt men wel de consistentie-eigenschap.

3. Foutvoortplanting

(3.1) We bestuderen nu de verschillen $\{e_n\}$, $e_n = X_n - X_n^*$, van een benaderde en een ware oplossing van een diff. verg. De algemene theorie hiervoor is niet eenvoudig. Wat er kan gebeuren wordt echter al heel aardig geïllustreerd door de simpele dvgl:

(3.2) $$X' = pX + q(t)$$

te beschouwen. Het is geen beperking $t_0 = 0$ te nemen, zodat $t_n = n.h$. We bekijken eens de methode behorend bij (2.4):

(3.3) $$X_n = X_{n-1} + h f_{n-1}$$

Hiervan (2.4) aftrekken geeft:

(3.4) $$e_n = h\delta_n + (1 + hp)\, e_{n-1}$$

(N.B $h\delta_n = h^2/_2\ x''(\xi)$.)

Schrijf $1 + hp = \sigma$. Als we de recurrente betrekking in (3.4) uitschrijven krijgen we:

(3.5) $$e_n = h\delta_n + \sigma(h\delta_{n-1} + e_{n-2}) =$$

$$= h\delta_n + \sigma h\delta_{n-1} + \sigma^2 h\delta_{n-2} + \cdots + \sigma^{n-1} h\delta_1 + \sigma^n e_0$$

Aangezien X_0 gegeven is geldt $e_0 = 0$.

Om hier een betere greep op te krijgen bekijken we wat er met σ^m gebeurt voor kleine waarden van h, waarbij we m niet onafhankelijk van h nemen. Wel geldt wegens $0 \leq t_m \leq T$ dat voor de ons interesserende waarden van m geldt $m.h \leq T$,

dus $\qquad m \leq T/h.$

Nu geldt

(3.6) $$\sigma = e^{\ln(1+ph)} = e^{ph + O(h^2)}$$

dus

(3.7) $$\sigma^m = e^{[p+O(h)]hm} = e^{[p+O(h)]t_m}$$

wegens $m.h \leq T$ is σ^m dus uniform begrensd in m en h (d.w.z er is een van m en h onafhankelijke bovengrens voor $|\sigma|^m$). Zij M een bovengrens voor σ^m. Voor (3.5) krijgen we dan:

(3.8) $$|e_n| \leq M.N.\max_{k=1,\dots,N}|h\,\delta_k| = M.T\max_{k=1,\dots,N}|\delta_k| = MT\cdot\frac{h}{2}\max(x''(\xi))$$

We zien zo dat e_n uniform naar 0 gaat, hetgeen een soort uniforme <u>convergentie</u> aangeeft van de benaderde oplossing naar de ware.

Een andere gevolgtrekking uit (3.5) is dat de bij de k^e tijdstap gemaakte fout $h\,\delta_k$ zich bij de n^e tijdstap als een σ^{n-k} maal zo grote fout manifesteert.

Of ook: de bij tijdstap $\tilde{t}$ gemaakte fout $h\,\delta_{\tilde{t}/h}$ manifesteert zich bij tijdstap t als een

$$e^{[p+O(h)](t-\tilde{t})} \quad \text{maal zo grote fout.}$$

Dit stemt treffend overeen met wat we zien als we de ware oplossing van de dvgl ten tijde $\tilde{t}$ een fout δ opleggen. We bekijken dan de oplossing $\tilde{x}$ met $\tilde{x}(\tilde{t}) = x^*(t) + \delta$.

Nu heeft het verschil van twee oplossingen* van (3.2) de gedaante $C.e^{pt}$ zodat:

$$\tilde{x}(t) - x^*(t) = \delta\, e^{p(t-\tilde{t})}.$$

De fout is dus met een factor $e^{p(t-\tilde{t})}$ aangegroeid; dit resultaat lijkt veel op het in (3.7) verkregene

(neem nl $t_0 = \tilde{t}$ en $t_m = t$). Globaal sprekend kunnen we dus zeggen dat bij (3.3) het maken van een (discretisatie- of afrond-) fout δ bij $\tilde{t}$ ongeveer neerkomt op het overstappen op de oplossing $\tilde{x}(\tilde{t})$ als boven, of ook dat fouten zich ongeveer voortplanten als oplossing van de homogene diff. vgl. Dit is in feite het gunstigste foutvoortplantingsgedrag dat men van een numeriek proces mag verwachten.

(3.9) Men kan aantonen dat dit gedrag bij andere (nette) methoden ook opgaat.

(3.10). Uit het voorgaande bleek dat het niet voldoende is dat de bij elke stap gemaakte (afrond- en discretisatie-) fouten klein zijn. Men moet ook eisen dat de eenmaal gemaakte fouten redelijk voortgeplant worden. Dit is een <u>stabiliteits</u>probleem.

(3.11). <u>Opgave</u>: Toon voor Euler backwards (2.5) uniforme convergentie aan, analoog aan het verhaal van Euler.

N.B. U moet vinden $\sigma = \frac{1}{1-hp}$.

(3.12) Zelfde vraag voor de trapeziumregel (2.6)

$$\left(\sigma = \frac{1+\frac{1}{2}hp}{1-\frac{1}{2}hp} \right).$$

4. Stijve differentiaalvergelijkingen

(4.1) Men zal wensen (en wellicht ook verwachten) dat wanneer de gevraagde oplossing van de dvgl een fraai langzaam variërend karakter heeft, de oplosmethode een grote stap toelaat.

Dit gebeurt echter niet altijd, zelfs niet met de methoden die in het voorafgaande bevredigend uit de bus kwamen.

(4.2) Als voorbeeld beschouwen we de dvgl.

(4.3)
$$0.001\, x' = -x + 1$$

zodat in de notatie van (3.2) geldt $p = -1000$, $q = 1000$.

De algemene gedaante van de oplossing hiervan is

(4.4)
$$x = C e^{-1000t} + 1,$$

C een willekeurige constante, zodat elke oplossing 1 als $\lim_{t \to \infty}$ heeft. De oplossingen gaan zelfs bijzonder snel naar 1 toe. Bijv. de oplossing met $x(0) = 0$, dus $C = -1$ is reeds voor $t = 0.01$ nauwelijks meer van 1 te onderscheiden.

We bekijken de methode van Euler (2.4). Hiervoor is $\sigma^n = (1+hp)^n = (1-1000h)^n$ (zie § 3). Dus zodra $h > 0.002$ groeit een eenmaal gemaakte fout onbegrensd aan, en is de methode zo al numeriek instabiel (d.w.z. t.o.v. afrondfouten). Maar ook zonder afrondfouten gaat het mis voor $h > 0.002$, want

(4.5)
$$x_n = 1 + C'(1-1000h)^n$$

(vul maar in) met $C' = -1$ als $x_0 = 0$.

Dus voor geen enkele waarde van t is een redelijke stapgrootte toegestaan, waarmee we opnieuw constateren (zie (3.10)) dat een kleine locale discretisatie fout (daarvan is met $x(0)=0$ zeker sprake voor $t > 0.01$) niet voldoende is maar dat hij ook nog redelijk voortgeplant moet worden.

(4.6) Veel beter gedragen (2.5) en (2.6) zich voor (4.2). Bijv.

bij (2.5) is $\sigma_n = \frac{1}{(1-hp)^n} < 1$ voor elke negatieve p en positieve h, dus gemaakte fouten sterven uit. Voor (2.5) en (4.2):

(4.7) $$x_n = 1 + \frac{C''}{(1+1000h)^n}$$

(4.8) De dvgl. (4.2) heeft de eigenschap dat de oplossingen der homogene vgl. snel dalen t.o.v. de bedoelde oplossing. Een dvgl. met deze eigenschap noemt men <u>stijf</u>. Op zich heeft zo'n dvgl. de prettige eigenschap dat verstoringen van de gewenste oplossing snel uitsterven (deze voldoen immers aan de homogene vgl), en voor voldoend kleine h is dat dan ook het geval met het effect van discretisatie fouten (zie (§ 3)).

Evenwel hangt het foutvoortplantingsgedrag af van ph; zie bijv. (3.4), en algemeen geldt voor (2.12) en (3.2)

$$(1-ph\beta_0)e_n + (\alpha_1 - ph\beta_1)e_{n-1} + \dots + (\alpha_k - ph\beta_k)e_{n-k} = 0.$$

Dus waar in (§ 3) sprake is van h klein genoeg betekent dit in feite dat ph klein genoeg moet zijn, zodat bij een stijve dvgl. h zeer klein moet zijn, en niet meer gerelateerd aan het fraaie gedrag van de oplossing.

We zien dan ook dat bij Euler en Euler backwards $\{\sigma^n\}$ zich helemaal niet meer ongeveer gedraagt als een oplossing van de homogene dvgl (zie de uitdrukkingen voor σ^n hierboven) wanneer h niet zeer klein is. Alleen bij Euler ontploft $\{\sigma^n\}$, terwijl hij bij Euler Backwards naar 0 blijft gaan en idem bij de trapezium regel.

(4.9) De bij stijve differentiaalvergelijkingen optredende situatie is ook heel goed grafisch te beschrijven. Beschouw de dvgl.

(4.10) $$\varepsilon x' = -x + g(t)$$

$0 < \varepsilon << 1$, g langzaam variërend. In fig 4.1 is het richtingsveld geschetst, en dat staat vrijwel

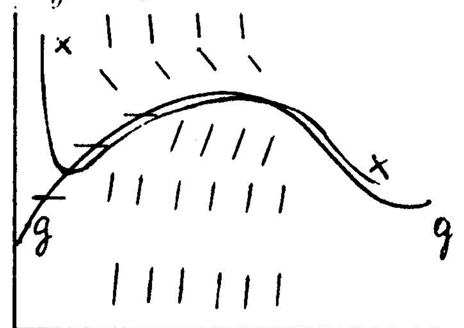


fig 4.1

overal verticaal, behalve op een onmiddellijke omgeving van de grafiek van g. Dus begeeft elke oplossing, waar ook gestart, zich onmiddellijk naar g, en blijft daar verder dicht in de buurt.

In fig 4.2 ziet men Euler weergegeven. Zodra men door de discretisatiefout iets te ver van de grafiek van

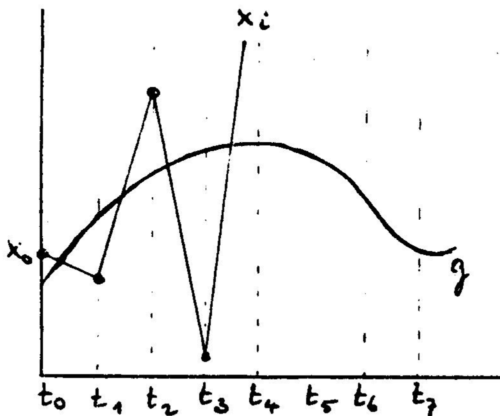


fig 4.2

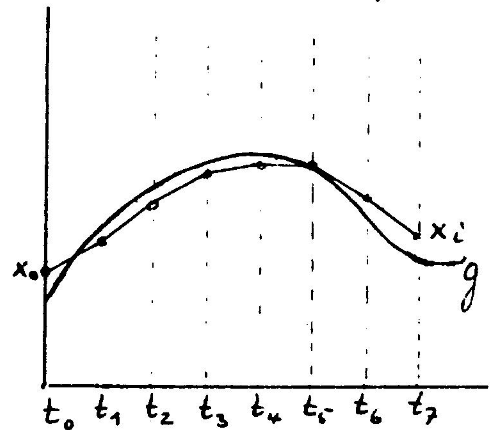


fig 4.3

g vandaan komt valt men in de klauwen van het verticale richtingsveld.

In fig 4.3 ziet men Euler backwards. Grafisch komt deze methode er op neer dat men bij gegeven x_0 een punt x_1 op de verticaal door t_1 moet zoeken waar het richtingslijnstukje door x_0 gaat, en het is duidelijk dat als x_0 dicht bij de grafiek van g ligt, x_1 niet ver van die grafiek vandaan kan liggen, juist wegens het verticale richtingsveld.

(4.11) Stijve differentiaalvergelijkingen komen veel vaker voor dan men wellicht denkt. Met name treden ze op zodra men bij het opstellen van de differentiaal vergelijking een klein effect laat meespelen zonder het welk de differentiaal vergelijking een lagere orde (of een stelsel een lagere dimensie) zou hebben gekregen. Het meenemen van zo'n effect noemt men wel het aanbrengen van een singuliere storing op de vergelijking (of het stelsel). Dit is heel duidelijk bij (4.10), waar men voor $\varepsilon = 0$ nagenoeg dezelfde oplossing krijgt als voor ε klein, maar waarbij dan de dvgl. van de eerste orde wordt.

(4.12) Ook bij niet-lineaire dvgln komen stijfheids-verschijnselen voor, en ook dan staan Euler backward en trapezium regel redelijke stapgroottes toe. Evenwel gaat de successieve substitutie methode uit § 2 niet meer op. Men gebruike nu (Koorden-) Newton of i.d.
(zie later)

(4.13) Naast de niet zo erg nauwkeurige (2.5) en (2.6) (lage orde!) zijn heel wat andere methoden bedacht om stijve dvgln. efficient op te lossen. Een behoorlijke theorie omtrent de vreemde verschijnselen die zich hierbij voor kunnen doen begint echter nu (1973) pas te ontstaan.